Evaluation and Health Status Prediction Method of Beer Filling Production Line Based on Data Mining Technology

Guo-cheng Niu and Zhen Hu

***Abstract*—**Considering the current situation that the system Health index of beer filling production line is hard to assess, quantify and predict, a prediction method of system health status based on the support vector machine(SVM) is proposed. Based on the principle of information entropy, through the big data analysis method, this paper quantitatively analyzes the behavior patterns and correlations between the internal attributes of the system, and calculates the real-time Health index of the production line system. The SVM method is used to predict the future bearing capacity of the production line, and the cross validation method and genetic algorithm are used to optimize the parameters (*c* and *g*) of SVM, and to construct the prediction model of Health index of filling production line. Finally, the simulation experiment is made to verify this method. The results show that this method is correct and feasible. By using this model, the prediction accuracy of the Health index of the filling production line can reach 0.9254, which can better guide production process improvement, equipment maintenance and production scheduling, and provide strategic support for scientific assessment and energy saving optimization of filling production line.

***Keywords*—**beer filling production line, information entropy, health index, support vector machine, health assessment.

This work was supported in part by Jilin Provincial Department of Education "13th Five-Year Plan" Science and Technology Project(JJKH20180338KJ)

Guocheng Niu is with the College of Electronic Information Engineering, Changchun University Science and Technology, Changchun 130022, Jilin, China.

Zhen Hu is with the College of Electronic Information Engineering, Changchun University Science and Technology, Changchun 130022, Jilin, China (corresponding author; e-mail: hanbingtbm@yeah.net).

I. INTRODUCTION

BEER filling is the main link of beer production. The system has a large scale and more and more equipment. There is a complex coupling relationship within the system, so its reliability is difficult to be guaranteed. Once the failure occurs, the loss of the equipment shutdown will be very large. The efficiency of the beer filling production line is difficult to have a reasonable and scientific evaluation and quantitative method, so it is more difficult to predict the health index of the production line in the future. At present, the commonly used method of beer enterprises is the calculation method of KPI parameters. Its rules are simple and their adaptability is not strong. In view of the above problems, this paper collects the real-time comprehensive information, such as the energy consumption, real-time production, liquor wastage and key performance indicator (KPI) in beer filling production. Based on the principle of information entropy, the calculation and quantification method of health index of filling production line is established. According to the historical Health index data, the SVM method is applied to predict the future health index of equipment to form a new method for evaluating and predicting the operation of the all-directional beer filling production line.

At present, scholars at home and abroad study on the health status assessment of complex systems from the point of view of monitoring data [1], review on fault prognostic methods based on uncertainty [2], a fault prognosis method using Bayesian network [3-4] and failure rate forecasting method based on neural networks. These methods are used to analyze the system health by modeling the system directly. If we want to comprehensively monitor the operation of large complex equipment and ensure efficiency and safety and maintainability, then maintenance strategy should be transformed from traditional abnormal state monitoring to health management [7]. For example, the U.S. military is proposed Prognostics and health management (PHM) [8], operational risk assessment based on health and importance

indexes for distribution network[9] and fault set classification method for power system reliability evaluation[10]. The core idea is to integrate data with intelligent algorithm based on the least number of sensor data, so as to realize the prediction and health monitoring of faults.

The information entropy theory is used to make the correlation mining for production data to calculate the heath degree of the production line. Data mining technology can show the unknown rules in a large amount of information, so it has more and more applications in the complex production environment. In complex production systems, there will be some fixed behavior patterns in the production link, equipment and operation state. All behavior patterns interweave together to form the running state and production capacity of the production system. Through data mining technology, we can accurately analyze the interaction relationship between behavior pattern attributes. Using reasonable association rules and scientific quantitative standards, the operation reliability and Health index of the system can be evaluated scientifically. This paper uses information entropy based correlation information to analyze and quantify the health index.

II. Calculation Theory of Health index

A. Data Selection

The sliding window model is used to collect data and obtain some new data for analysis. All the data to be processed is expressed in the form of a two-dimensional matrix with attributes and values as coordinates.

B. The Construction of Pattern Diagram and the Selection of Evaluation Precision

The behavior patterns of variables in the data should be determined. In N variables, two variables are taken as the X axis (the base attribute) and the Y axis (the reference attribute) and they are taken as a behavior pattern. N (N-1) behavior patterns can be obtained by pairwise permutation and combination. The collected continuous variables are discretized by means of constant width segmentation method. The constant width segmentation area is the sub pattern of the behavior patterns. The interval number (EP) (value is 5-9) is the assessment accuracy. The associated information of all subpatterns is added together, and it is taken as the associated information of this behavior pattern. The associated information of all patterns is integrated together, and it is taken as the running properties of the system at this time.

C. Association Rules and Minimum Confidence

When calculating the associated information, we need to cal culate the basic quantity of quantity—minimum confidence an d minimum support. It reflects the extent to which the Y axis p roperty is disturbed when the X axis property changes. Satisfyi ng minimum confidence indicates that the system has enough possibility to fall into this state during the operation. Satisfying the minimum support indicates that when it is disturbed, the sy stem state has enough possibility to change to another state aft er being disturbed.

Calculation formula is min_con=2/EP min_SUP=2/EP.

D. Quantification of Evaluation Attributes

If you want to know the strength of the association, you need to quantify the association rules. Generally, entropy is used to describe the degree of disorder of the rule of the system. In this paper, the definition of the strong and weak entropy which is used to describe the association rules is as follows,

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_s p(x_i) \tag{1}$$

Among them, $p(x_i)$ is the probability that the X value is equal to x_i , i=1,2,…, n. The unit of the information entropy is related to the s in (1). Here s=e is taken, then the information entropy unit is nat.

Mutual information, as a measure of the degree of association between two variables, is defined as

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{2}$$

Finally, the generalized correlation function R_g is used to quantify the correlation relationship, and describe the strength and weakness of the association property. The value of R_g is (0-1). The stronger the correlation between X and Y, the closer the R_g to 1.

$$R_g = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \tag{3}$$

E. Synthesis of System Health index

For a certain pattern, its Health index is the sum of the products of each sub pattern and confidence, that is,

$$U(A_x, A_y) = \sum_{i=1}^{EP}\sum_{j=1}^{EP} confidence(i,j)R_{i,j}\left|C_{A_x,A_y}\right. \tag{4}$$

Among them $R_{i,j}$ is the generalized correlation coefficient of the cell (i, j) mining the subpattern area. C_{A_x,A_y} indicates that this calculation is located in the coordinate system of A_x and A_y.

The health index of the whole system is the association relationship of all behavior patterns of all variables pairwise permutation, that is the set of the Health index. The systems of the N variables can get N*(N-1) patterns at most. The formula of system health index is:

$$U = \sum_{x=1,y=1,y\neq x}^{n} \sum^{n} U\left(A_x, A_y\right) = \sum_{x=1,y=1,y\neq x}^{n} \sum^{n} \left(\sum_{i=1}^{EP} \sum_{i=1}^{EP} confidence(i,j) R_{i,j} \middle| C_{A_x,A_y} \right) \quad (5)$$

F. Rationality Analysis of the Calculation Method of Health Index

The key performance indicators (KPI) of the production Health index of traditional assessment filling line are the total asset utilization ratio, the line gross output rate, the total equipment utilization rate and the line efficiency. The effects of energy consumption and beer consumption on the production Health index are not considered, and the indexes are unscientific and incomplete. The evaluation method in this paper takes the 7 variables of unit energy consumption, liquor wastage, unit output and 4 KPI parameters as the mining reference variables. Energy consumption, liquor wastage and output data are derived from the energy management system, and KPI data is derived from the control system of the filling workshop. The 7 variable data curve of the first filling line of a beer production plant in the 30 days of June 2016 is shown in Fig1.

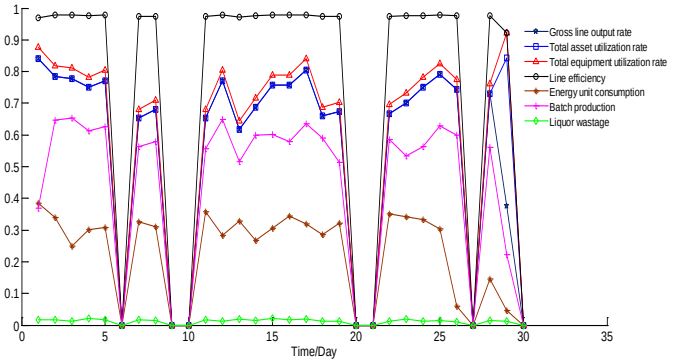

Fig. 1. The seven variable data samples for 30 days in June 2016

After normalizing the data samples, 42 pattern diagrams about 7 variables are set up, and the accuracy is 7. The original Health index of the system is calculated by formula (5). In actual production, the lower the energy consumption and the liquor wastage, the higher the efficiency. And the higher the other variables, the higher the efficiency of the production line. Therefore, after normalizing the variable, reverse processing is carried out for 5 variables except energy consumption and liquor wastage. To verify the practicability of the method, one of the variables is adjusted each time. The Health index after the adjustment of the system data is calculated by the formula (5), and the results are shown in Table 1.

Table 1 Comparison table of change of equipment Health index

	Health index of the original month	Health index after adjusting parameters	Percentage of changes of the Health index
Total asset utilization ratio (Raise 50%)	70.813	73.129	3.27%
Line gross output rate (Raise 50%)	70.813	72.702	2.67%
Total equipment utilization rate (Raise 50%)	70.813	73.299	3.51%
Line efficiency (Raise 50%)	70.813	80.762	14.05%
Output (Raise50%)	70.813	71.471	0.93%
Energy consumption (Reduce 50%)	70.813	74.964	5.86%
Liquor wastage (Reduce 50%)	70.813	81.220	14.70%

From Table 1, we can see that the liquor wastage has the greatest impact on the Health index of the production line, and in KPI parameters, the line efficiency has the greatest impact on the Health index, and its change rule is consistent with the actual production assessment index. Using this method, we calculate the daily Health index of 7 variables in the new production line and old production line of a factory filling workshop in June and November 2016 by hourly sampling interval. The operation Health index of new and old production lines in June is shown in Fig.2. In June, the average Health index was 39.877 and 26.788 respectively. The operation Health index of new and old production lines in November is shown in Fig.3. In November, the average Health index was 20.23 and 17.01 respectively.

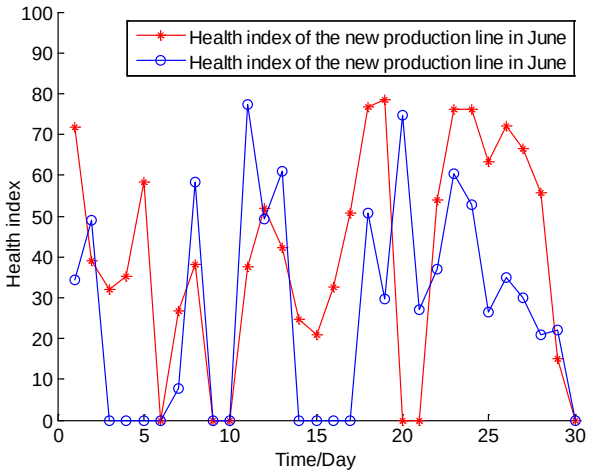

Fig .2. The Health index of the two production lines in June

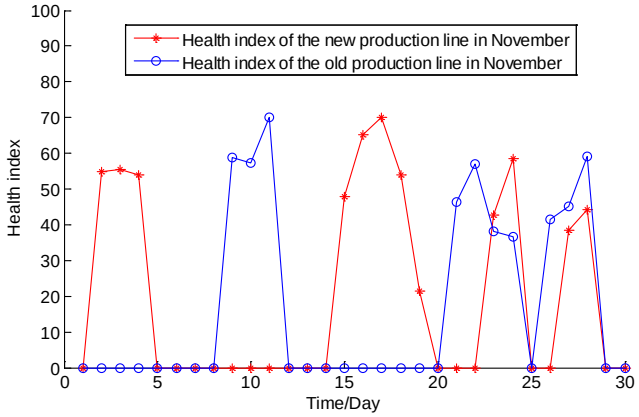

Fig. 3. The Health index of the two production lines in November

According to Fig.2 and 3, in November, the new production equipment had large output, less downtime, less liquor wastage, good heat preservation and less energy consumption. June is the peak production season. The equipment continuous running time is long, the ambient temperature is high, the KPI parameter is good, and the production efficiency is high. November is the off-season, so the production is intermittent. The Health index of the production line in June was better than that in November. The comparison between theoretical analysis and actual production condition shows that this set of analysis plan is reasonable and accurate, and the Health index of production line is quantified, which has certain reference value.

III. Health Index Prediction of Production Condition

In production, monitoring the Health index of equipment in real time is very important to the production guidance. However, managers prefer to use historical production efficiency information to predict the equipment health status of the next production cycle. It will play a greater role in production scheduling, equipment maintenance and energy conservation. In this paper, a support vector machine model with parameter optimization is used to predict the future health status of the production line.

A. SVM Prediction Method

Support vector machine (SVM) is an approximate implementation of structural risk minimization. It has good nonlinear mapping ability and can overcome the inherent problem of neural network "dimension disaster". It has good application effect under the condition of small sample, and its main idea is:

Give l independent identically distributed training sample $(x_1, y_1),(x_2, y_2),\cdots(x_l, y_l)$, $i = 1,2,\cdots,l$, $x_i \in R^m$ is the input of the m-dimensional training sample, $y_i \in R$ is the output of the training sample. SVM seeks an optimal function

$$f(x) = (x, \omega) + b, \omega \in R^n, b \in R \tag{6}$$

The optimal decision function is obtained:

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*)(x, x_i) + b \tag{7}$$

In formula (7), α_i, α_i^*, b quadratic programming problem is obtained, that is the dual problem of solving the optimization problem.

$$\begin{aligned} \min Q(\alpha_i, \alpha_i^*) = \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(x_i, x_j) \\ -\sum_{i=1}^{l}(\alpha_i + \alpha_i^*)y_i + \sum_{i=1}^{l}(\alpha_i + \alpha_i^*)\varepsilon \end{aligned} \tag{8}$$

In formula (8), ε is an insensitive loss function that represents the absolute error limit of the predictive value of the support vector machine.

The constraint condition is as follows

$$\begin{aligned} &\sum_{i=1}^{l}(\alpha_i - \alpha_i^*)y_i = 0 \\ &0 \le \alpha_i \le C \\ &0 \le \alpha_i^* \le C \end{aligned} \tag{9}$$

By solving the above problems, the fitting function is finally obtained.

$$f(x) = \sum_{i=1}^{n} W_i K(x, x_i) + b \tag{10}$$

In formula (10), $K(x, x_i) = (x, x_i)$ is a kernel function that

satisfies the Mercer condition, $W_i = (\alpha_i - \alpha_i^*)$ is the support vector coefficient, x_i is the support vector, x is the sample to be predicted, n is the number of support vector, b is the constant.

B. Health index Prediction of Production Status

In the Matlab 7.11.0 development environment, time is taken as input, Health index is taken as output. According to the formula (5), the data of filling production line of June in 2014-2016(except the shutdown period) is used to calculate the Health index. We take the sample once an hour and calculate the Health index of the effective production date by day. There are 65 sets of sample data, which are divided into training set (40 groups) and test set (25 groups). The radial basis function (RBF) kernel function

$$K(x, x_i) = \exp\left(-g\|x - x_i\|^2\right) (g > 0) \tag{11}$$

It is used as the kernel function of SVM. Cross validation's grid-search and genetic algorithm are used to optimize parameters (penalty parameter c and RBF function's span coefficient g), which are respectively called K-SVM and GA-SVM prediction methods. The decision function of SVM is used as the predictive function of the Health index. The fitting curve of the training set is shown in Fig.4.

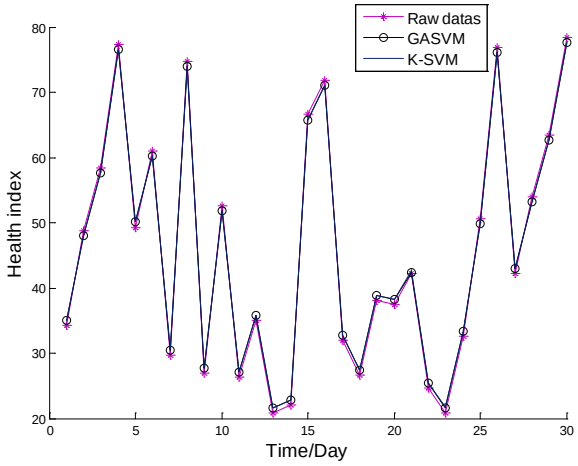

Fig .4. SVM model fitting curve

C. Verification Results and Performance Analysis

In order to verify the practicability of the SVM algorithm for Health index prediction, two decision functions of the K-SVM and GA-SVM algorithms obtained above are used to predict the test set. The fitting curve of the test set is shown in Fig.5.

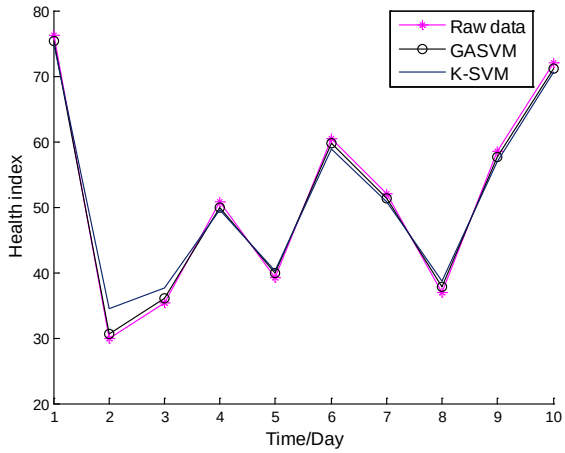

Fig .5. Test set fitting curve graph

After repeated simulation experiments, the performance indexes of SVM modeling of two optimization methods listed in Table 2 are respectively the optimal parameters (c and g), normalized mean square error (MSE) and correlation coefficient R.

Table. 2. Performance comparison of two methods

Performance	Method	
	K-SVM	GA-SVM
Optimal parameters	*c*= 13.4543 *g*=0.25	*c*= 31.4553 *g* = 1.9159
MSE	0.1287	0.0534
R	82.45%	0.9254

From Fig.5 and Table 2, it can be known that when the GA-SVM prediction method is used, c= 31.4553 and g = 1.9159, the MSE of the test set is 0.0534, and the autocorrelation coefficient is 0.9254, indicating that the prediction accuracy of this model for the Health index of the filling production line reaches 92.54%, and it meets the requirements of the actual production guidance precision.

IV. Conclusion

According to the actual demand of beer filling production line, the energy consumption, product yield, liquor wastage and KPI operation parameters in beer filling production are used as the raw data to measure the Health index of the system. The information entropy principle is used to evaluate the behavior patterns of many variables and the degree of attribute correlation in the information. The Health index of the quantified system is analyzed statistically, and the results are in conformity with the actual production. According to the historical Health index, the SVM algorithm is used to predict the health status of the production line in the next production cycle, achieve the prediction of the operation capacity of the production line, and guide the process improvement, equipment maintenance and production scheduling.

References

[1] Ningyun LU, Kelei HE, and Jianu Bin. "A fault prognosis method using Bayesian network", *Journal of Southeast University: Natural Science Edition*, vol.42 , no.15, pp. 87-91, 2012.

[2] Qiang SUN, Jiguang YUE. "Review on fault prognostic methods based on uncertainty", *Control and Decision*, vol.29, no.5, pp. 769-778, 2014.

[3] Ruiying LI, Rui KANG. " Failure rate forecasting method based on neural networks" *Acta Aeronautica Sinica*, vol.29, no.2, pp. 357-363, 2008.

[4] ALI J B, FNAIEC'H N, and SAIDI I. "Application of empirical mode decomposition and artificial neural network for automatic bearing fault diagnosis based on vibration signals", *Applied Acoustics*, vol.201, no.89, pp. 16-27, 2011.

[5] Colantonio S, Di Bono M G, and Pieri G. "System health state monitoring using multilevel artificial neural networks", *Computational Intelligence for Measurement Systems and Applications*, vol.21, no.8, pp. 50-55, 2005.

[6] Junxun CHEN, Longsheng CHENG, and Hui YU. "Health status assessment for complex systems based on EMD-SVD and Mahlanobis-Taguchi System", *Systems Engineering and Electronics*, Vol.39 , No.7, pp. 1542-1548, 2017.

[7] Boyuan LIU, Huangang WANG, and Wenhui FAN. "Real time health level assessment for complex production line system based on big data", *Tsinghua Univ*, vol.54, no.10, pp. 1377-1383, 2014.

[8] Hess A, Fila L. "The joint strike fighter (JSF) PHM concept: Potential impact on aging aircraft problems", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systms*, vol. 6, no.10, pp. 3021-3026, 2002.

[9] Zhao HUANG, Feng WANG, and Yanghong TAN. "Operationalrisk assessment based on healthand importance indexes fordistribution network", *Electric Power Automation Equipment*, vol.36, no.6, pp.136-141, 2016.

[10] Jiangning HUANG, Ruipeng GUO, and Fang ZHAO, "Fault set classification method for power system reliability evaluation", *IEEE Trans, on Industrial Electronics*, vol.33, no.16, pp. 112-119, 2013.

[11] ALI M, AHN C W, and PANT M. "A robust image watermarking technique using SVD and differential evolution in DCT domain", *Optik-International Journal for Light and Electron Optics*, vol.125, no.1, pp.428-434, 2014.

[12] Jisheng XING, and Haiwei WU. "Prediction model of energy consumption on beer enterprise based on support vector machine", *Journal of Jilin University*, vol.32, no.6, pp.664-669, 2014.

[13] Balakrishnan K J, and Touba N A. "Relationship between entropy and test data compression", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systms*, vol.26, no.2, pp. 386-395, 2007.

[14] Colantonio S, Di Bono M G, and Pieri G. "System health state monitoring using multilevel artificial neural networks", *Computational Intelligence for Measurement Systems and Applications*, vol.26, no.2, pp. 50-55, 2015.

[15] CHEN Junxun, CHENG Longsheng, and YU Hui, "Health status assessment for complex systems based on EMD-SVD and Mahlanobis-Taguchi System", *Systems Engineering and Electronics*, vol.39, no.7, pp.1542-1548, 2017.

[16] ALI J B, FNAIEC'H N, and SAIDI I. "Application of empirical mode decomposition and artificial neural network for automatic bearing fault diagnosis based on vibration signals", *Applied Acoustics*, vol.89, no.3, pp. 16-27, 2015.

[17] ALI M, AHN C W, and PANT M. "A robust image watermarking technique using SVD and differential evolution in DCT domain", *Optik-International Journal for Light and Electron Optics*, vol.125, no.1, pp. 428-434, 2014.

[18] JAVED K, UOURIVEAU R, and ZERHOUNI N. "Enabling health monitoring approach based on vibration data for accurate prognostics", *IEEE Trans, on Industrial Electronics*, vol.62, no.1, pp. 647-656, 2014.

[19] JIANG Chcng, LIU Wcnxia, and ZHANG Jianhua. "Risk assessment of generation and transmission systems considering wind power penetration", *Transactions of China Elcctrotcchnical Society*, vol.29, no.2, pp.260-270, 2014.

[20] ALVEHAG K, and SODER L. "Risk-based method for distribution system reliability investmentdecisions under pcrlormancc-based rcgulation", *IET Gencration,Transmission & Distribution*, vol.5, no.10, pp.1062-1072, 2011.

Guocheng Niu was born on May. 2, 1977**.** He's a doctoral graduate student at the Changchun University of Science and Technology. Currently, he is an associate professor at Beihua University, China. The main research direction is intelligent information processing. He has published many papers in the relevant periodicals.

Zhen Hu was born on Nov. 24, 1962**.** He received a Ph. D. in electronic information engineering at Changchun University of Science and Technology in China. At present, he is a professor at Hunan University, China. The main research direction is intelligent information processing and Internet of things. He has published many papers in the relevant periodicals.